


# Improving Credit Card Fraud Detection in Imbalanced Datasets: A Comparative Study of Machine Learning Algorithms

Ruozhang Liu <sup>a</sup>

*D'Amore-McKim School of Business, Northeastern University, 360 Huntington Ave, Boston, U.S.A.*

**Keywords:** Credit Fraud, SMOTE Analysis, SVM Analysis.

**Abstract:** This research focuses on tackling the challenge of identifying credit card fraud within highly imbalanced datasets, where the proportion of fraudulent transactions is significantly smaller compared to the overall number of transactions. Using a dataset from the Kaggle, this study applied various preprocessing techniques, including normalization, data cleaning and undersampling and so on to balance the data. This paper aims to evaluate several machine learning algorithms—Logistic Regression, K-Nearest Neighbors, Support Vector Machines (SVM), and Decision Trees—based on different metrics. Logistic Regression and SVM showed the best performance, balancing precision and recall effectively. Despite improvements, the trade-off between precision and recall remains a challenge, indicating the need for more advanced methods like ensemble learning and deep learning. The findings emphasize the importance of sophisticated machine learning techniques in improving the accuracy and reliability of credit card fraud detection systems, ultimately protecting financial institutions and customers from significant financial losses.

## 1 INTRODUCTION


The importance of business analytics and machine learning in today's fast-moving business environment cannot be overemphasized. These advanced technologies have become an integral part of the modern business environment, driving data-driven decision-making processes that result in increased operational efficiencies, improved customer experience, and increased profitability.

Despite the broad applicability of business analytics and machine learning across various industries, their specific use in the financial sector, especially in detecting credit card fraud, has garnered considerable interest. Credit card fraud is a type of identity theft and financial crime that involves the illicit use of someone else's credit card details to make transactions, withdraw money, or gain other financial advantages. This illegal activity can occur in a variety of forms, such as cardless fraud, skimming, and phishing. The financial implications of credit fraud are far-reaching and can result in significant losses to individuals and financial institutions.

The challenge of detecting credit card fraud is exacerbated by the fact that the datasets involved are

highly imbalanced, with fraudulent transactions accounting for only a small fraction of total transactions. As a result of this imbalance, traditional methods are often ineffective at identifying fraudulent activity, leading to high rates of false positives or underreporting. This requires more sophisticated machine learning techniques and data processing methods to improve the accuracy and reliability of fraud detection systems (Zou, 2018; Zheng et al., 2020; Ngai et al., 2011).

This research focuses on solving the data imbalance problem in credit card fraud detection by applying machine learning algorithms and advanced sampling techniques. Two main strategies are employed: undersampling and oversampling. These approaches aim to create a balanced dataset to improve the performance of machine learning models. In this context, the study evaluated various machine learning models, including logistic regression, decision trees and random forests etc., to determine the most effective fraud detection methods. Each of these algorithms has unique strengths (Kotsiantis, 2006): logistic regression provides a probabilistic framework, decision trees provide interpretability, random forests add robustness, and

<sup>a</sup> <https://orcid.org/0009-0007-4085-1810>

SVMs ensure effective classification in high-dimensional spaces. This research seeks to build a robust and reliable fraud detection system by leveraging the strengths of different machine learning algorithms and tackling data imbalances, aiming to safeguard financial institutions and their customers from the detrimental impacts of credit fraud.

## 2 METHOD

### 2.1 Dataset Preparation

#### 2.1.1 The Introduction of the Dataset

The dataset utilized in this study, sourced from Kaggle by Janio Martinez Bachmann (Janio, 2019). Spanning over two days, the dataset includes 284,807 transactions, of which 492 are fraudulent, highlighting its highly imbalanced nature. Due to confidentiality concerns, the dataset features have undergone Principal Component Analysis (PCA) transformation, with only the transformed features being available and the original features omitted. It consists of 31 columns, including transaction time, transaction amount, the principal components labeled V1 through V28, and a binary label identifying whether a transaction was fraudulent (1) or legitimate (0).

### 2.2 Explanation of the Class Distributions Plot

The provided Figure 1 represents the distribution of the classes.

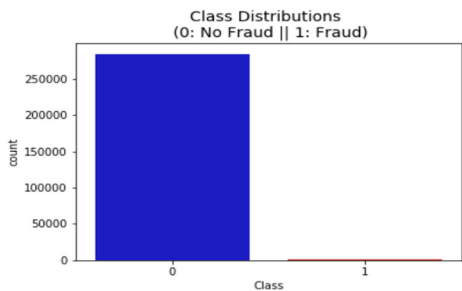


Figure 1: Class Distribution (Photo/Picture credit: Original).

The plot vividly illustrates the severe imbalance in the dataset. The blue bar represents the count of non-fraudulent transactions (Class 0), which significantly outnumbers the fraudulent transactions (Class 1) represented by the small red bar.

The blue bar shows that there are over 250,000 non-fraudulent transactions in the dataset. This indicates that the dataset is heavily skewed towards non-fraudulent transactions. The red bar, barely visible in comparison, shows that there are fewer than 1,000 fraudulent transactions.

#### 2.2.1 Explanation of the Distribution of Transaction Amount and Time

**Distribution of Transaction Amount (Left Plot):** The left plot shown in Figure 2 shows the distribution of transaction amounts, which is highly right-skewed. Most transactions have relatively small amounts, with the majority clustered around lower values and a steep drop-off as the amount increases. A few transactions have very high amounts, but these are rare and fall far to the right of the plot.

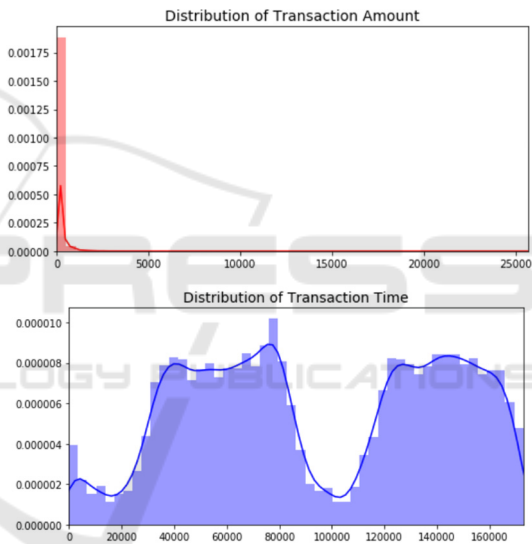


Figure 2: Distribution of Transaction Amount and Transaction Time (Photo/Picture credit: Original).

**Distribution of Transaction Time (Right Plot):** The right plot displays the distribution of transaction times. The distribution shows multiple peaks and troughs, suggesting periodic patterns or cycles in transaction activity over time. This indicates that there are specific intervals with higher transaction activity.

#### 2.2.2 Explanation of Equally Distribution Classes

For credit card fraud detection, the original dataset is significantly imbalanced, with fraudulent transactions representing only a small fraction of the total. This imbalance can adversely affect model

performance, as the model may become skewed towards the majority class (non-fraudulent transactions), potentially leading to reduced effectiveness in detecting fraudulent activity. By balancing the dataset, either through under sampling the majority class or oversampling the minority class (e.g., using SMOTE), it can be ensured that the model has a better chance of learning the characteristics of both classes effectively.

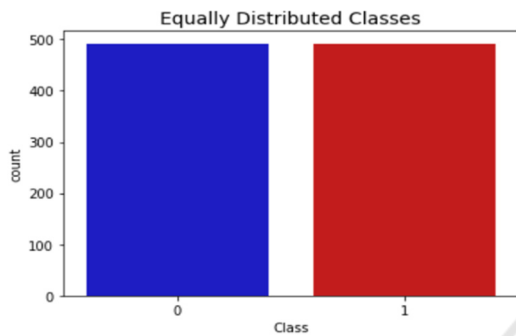


Figure 3: Equally Distribution Classes (Photo/Picture credit: Original).

This visualization shown in Figure 3 illustrates the process of balancing the dataset, which is a crucial step in improving the performance of machine learning models for fraud detection. By ensuring an equal distribution of classes, the model can learn to distinguish between fraudulent and non-fraudulent transactions more accurately, leading to better detection rates and fewer false positives/negatives.

### 2.2.3 Preprocessing Steps Based on Distribution Plots

The distribution plots for the V14, V12, and V10 features for fraudulent transactions provide valuable insights that inform the preprocessing steps. These steps include normalization, data cleaning, balancing the dataset, applying SMOTE, and splitting the data into training and testing sets.

**Normalization:** Normalization is crucial to ensure that all features have an equal influence during model training, preventing any single feature from dominating the learning process due to differences in scale. The skewed nature of some features, as observed in the distribution plots, indicates the need for scaling.

**Cleaning:** Data cleaning involves handling missing values, duplicates, and irrelevant features. In the dataset, this study assumes no missing values based on initial observations, but this study drops irrelevant columns after transformation.

**Balancing the Dataset:** Given the highly imbalanced nature of the dataset, balancing is crucial. This study used undersampling and SMOTE to address this issue. Undersampling reduces the number of non-fraudulent transactions, while SMOTE generates synthetic samples for the minority class.

**Train-Test Split:** To evaluate the model's performance, this paper splits the dataset into training and testing sets. An 80-20 split ratio is commonly used.

## 2.3 Machine Learning-Based Prediction

### 2.3.1 Introduction of the Machine Learning Workflow

The machine learning workflow is a structured process for building models that yield accurate predictions. It begins with data collection, followed by data preprocessing, which involves cleaning, normalization, and addressing data imbalances. After preprocessing, the data is split into training and testing sets. Different machine learning algorithms are then applied to the training set to create predictive models, which are subsequently evaluated on the testing set. This structured approach ensures robust and reliable machine learning solutions (Witten et al., 2016).

**Data Collection:** Data collection is the initial phase of the machine learning workflow, involving the aggregation of relevant data from various sources. The quality and quantity of the collected data are critical, as they directly influence the model's effectiveness and overall performance. This data can be structured or unstructured and is typically gathered from databases, APIs, web scraping, or manual entry. High-quality data collection ensures that the subsequent steps in the workflow are built on a solid foundation (Zhang et al., 2019).

**Model Building:** Model building involves selecting the appropriate machine learning algorithms that will be used to create the predictive model. This step includes defining the model architecture, choosing the type of model (e.g., regression, classification), and setting hyperparameters. Common algorithms used in fraud detection include Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVM) (Zhou et al., 2018).

**Model Training:** Model training involves feeding preprocessed data into the chosen algorithm, enabling the model to learn patterns from the data. During this phase, the model adjusts its parameters to minimize

prediction errors through an iterative process, often involving multiple epochs or passes over the data. Each iteration helps the model progressively enhance its accuracy. To ensure the model generalizes effectively to new, unseen data, the training data must adequately represent the problem space (Goodfellow et al., 2016).

**Testing:** Testing involves assessing the trained model using a separate dataset that was not involved in the training process. Testing helps to detect any overfitting or underfitting, ensuring that the model is robust, reliable, and ready for deployment.

## 2.4 SVM-Based Prediction

### 2.4.1 Principle of SVM

SVM applies a kernel function to map the original data into a higher-dimensional space, making it easier to identify a linear separation between the classes. The resulting optimization problem is solved using convex optimization techniques, ensuring a globally optimal solution. SVM is highly effective in high-dimensional spaces and demonstrates strong resistance to overfitting (Ben-Hur & Weston, 2010).

### 2.4.2 Hyperparameters and Evaluation Metrics

Hyperparameters are the settings determined before the training process starts, governing the overall behavior and structure of the model's learning process. Unlike model parameters, which are derived from the data during training, hyperparameters are predefined by the user and adjusted manually to optimize the model's performance. Examples include the regularization parameter ( $C$ ) in SVM, the number of trees in a Random Forest ( $n_{\text{estimators}}$ ), and the kernel type in SVM ( $\text{kernel}$ ). Tuning hyperparameters is crucial for optimizing model performance.

## 3 RESULTS & DISCUSSION

This study applied various machine learning techniques to detect credit card fraud, addressing the significant challenge of dataset imbalance. The methods included data preprocessing, model training with different algorithms, and evaluation using appropriate metrics.

### 3.1 Data Preprocessing

The Amount and Time features were normalized using StandardScaler. The dataset was also examined for missing values, which were cleaned to maintain data integrity. Due to the extreme imbalance, with fraudulent transactions accounting for only 0.172% of the dataset, both undersampling and the Synthetic Minority Over-sampling Technique (SMOTE) were applied. Undersampling reduced the number of non-fraudulent transactions, while SMOTE generated synthetic samples for the minority class, helping to create a more balanced dataset.

### 3.2 Discussion

The application of SMOTE has enhanced the model's performance as evidenced by the high recall score in Figure 4, Figure 5 and Figure 6. However, the model's low precision for fraud indicates a potential issue with generating a large number of false positives. This trade-off between precision and recall is crucial in fraud detection. While high recall ensures the detection of most fraud cases, low precision can flood the system with false alarms, leading to unnecessary investigations and potential inconveniences for both the institution and its customers. Balancing precision and recall are essential to maintain an effective and efficient fraud detection system.

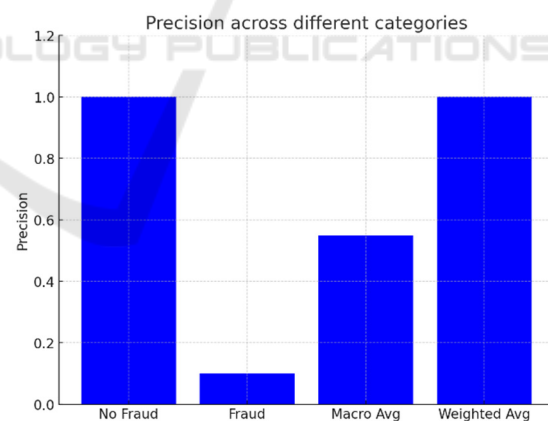


Figure 4: The precision results (Photo/Picture credit: Original).

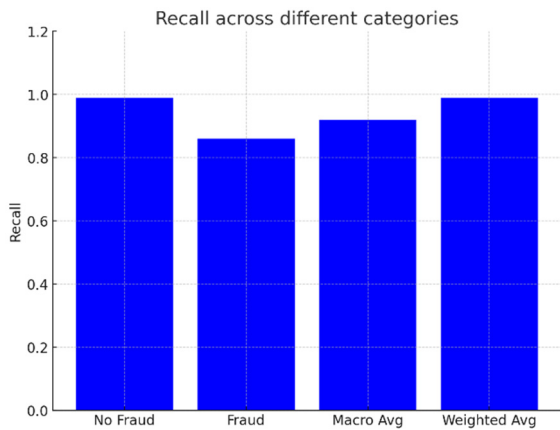


Figure 5: The recall results (Photo/Picture credit: Original).

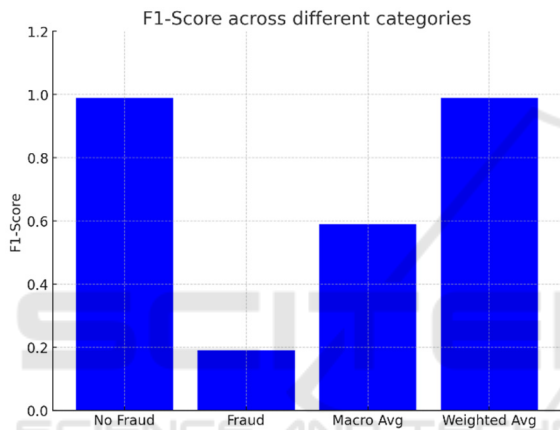


Figure 6: The F1-score results (Photo/Picture credit: Original).

Handling the imbalanced dataset was a significant challenge. While SMOTE improved recall by increasing the detection of fraudulent transactions, it sometimes led to lower precision, causing more false positives. This trade-off between precision and recall is critical in fraud detection, where the cost of false positives (unnecessary investigations) and false negatives (missed frauds) must be balanced.

Future research could delve into more advanced ensemble methods. Additionally, integrating deep learning techniques could further boost the model's capacity to identify intricate patterns within the data. Continuous monitoring and updating of models remain vital for adapting to the constantly evolving tactics of fraudsters.

## 4 CONCLUSIONS

This study underscores the significance of advanced machine learning techniques in credit card fraud detection. By implementing normalization, data cleaning, and balancing strategies like undersampling and SMOTE, a robust dataset was constructed for model training. The study evaluated several machine learning models. Logistic Regression and SVM proved to be the top performers, striking a balanced trade-off in performance. However, challenges persist, particularly with precision, indicating the need for further exploration of more advanced ensemble methods and deep learning approaches.

## REFERENCES

- Ben-Hur, A., & Weston, J. 2010. A user's guide to support vector machines. In *Data mining techniques for the life sciences* (pp. 223-239). Humana Press.
- Goodfellow, I., Bengio, Y., & Courville, A. 2016. *Deep Learning*. MIT Press.
- Janio, M. B. 2019. Credit Fraud. Dealing with Imbalanced Datasets. Kaggle.
- Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. 2006. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3), 159-190.
- Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559-569.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.
- Zhang, Z., He, S., Chen, X., & Liu, G. 2019. A machine learning-based framework for process data analysis. *Journal of Process Control*, 79, 112-122.
- Zheng, D., Wu, S., & Chen, H. 2020. Fraud detection in credit card transactions with SMOTE and ensemble learning. *Journal of Information Processing Systems*, 16(5), 1053-1064.
- Zhou, Y., Cheng, G., Jiang, S., & Wang, Y. 2018. A deep learning approach for credit card fraud detection. *Journal of Financial Crime*, 25(4), 1065-1079.
- Zou, W. 2018. An improved credit card fraud detection model based on SVM and LightGBM. *Journal of Financial Risk Management*, 7(4), 456-467.